To complete capstone, you will be working on a case study which is to predict the severity of an accident.

Now, wouldn't it be great if there is something in place that could warn you, given the weather and the road conditions about the possibility of you getting into a car accident and how severe it would be, so that you would drive more carefully or even change your travel if you are able to.

Let's use our shared data for Seattle city as an example of how to deal with the accidents data.

The first column colored in yellow is the labeled data.

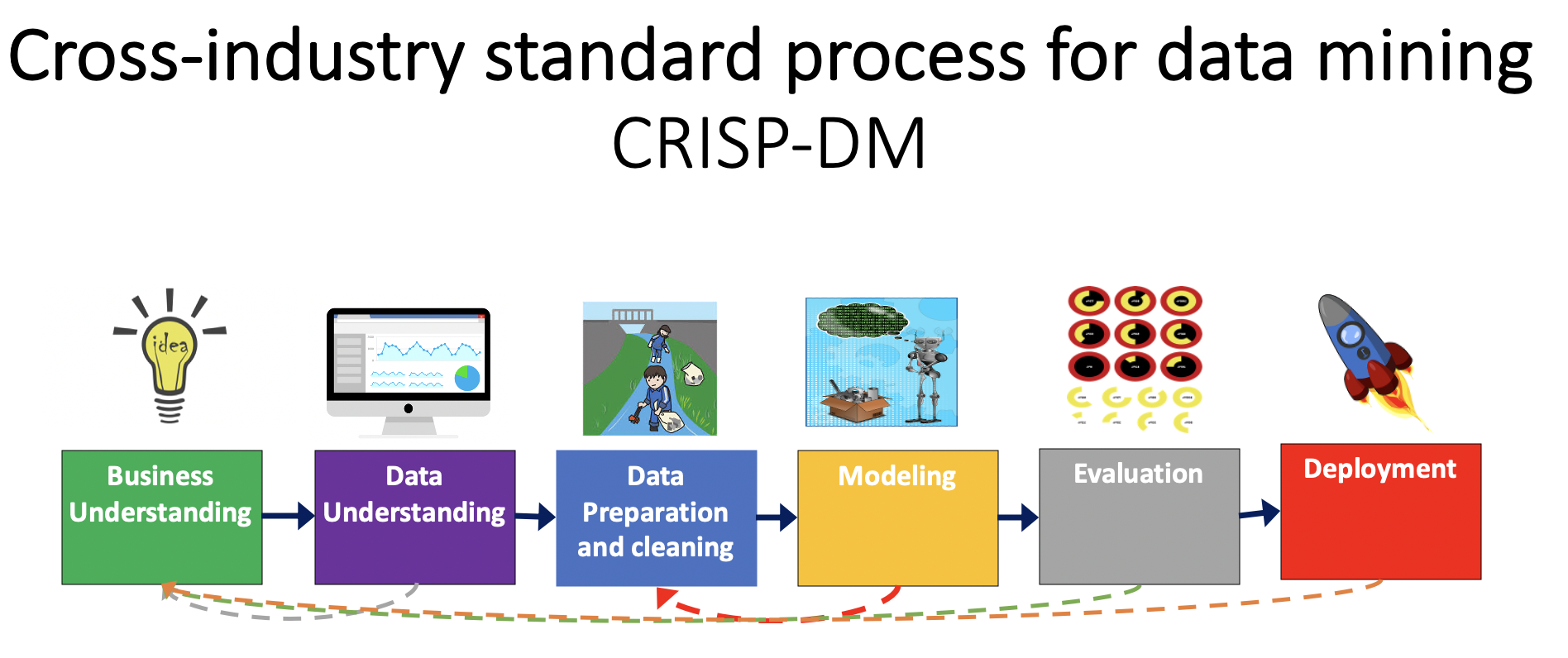
The label for the data set is severity, which describes the fatality of an accident.

You will notice that the shared data has unbalanced labels. You should balance the data, otherwise, you will create a biased ML model.

The remaining columns have different types of attributes. Some or all can be used to train the model. You can also find that most of the observations are good to train and test the machine learning model.

You might need to do some feature engineering to improve the predictability of your model.

1. The target or label columns should be accident " severity" in terms of human fatality, traffic delay, property damage, or any other type of accident bad impact.
2. The machine learning model should be able to predict accident "severity"
3. To build a good model, the dataset should be rich and contain many observations (rows) and various attributes (columns)



**Business Understanding:**

The initial phase is to understand the project's objective from the business or application perspective. Then, you need to translate this knowledge into a machine learning problem with a preliminary plan to achieve the objectives.

**Data understanding:**

In this phase, you need to collect or extract the dataset from various sources such as csv file or SQL database. Then, you need to determine the attributes (columns) that you will use to train your machine learning model. Also, you will assess the condition of chosen attributes by looking for trends, certain patterns, skewed information, correlations, and so on.

**Data Preparation:**

The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation can be performed multiple times and it includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

**Modeling:**

In this phase, various algorithms and methods can be selected and applied to build the model including supervised machine learning techniques. You can select SVM, XGBoost, decision tree, or any other techniques. You can select a single or multiple machine learning models for the same data mining problem. At this phase, stepping back to the data preparation phase is often required.

**Evaluation**:

Before proceeding to the deployment stage, the model needs to be evaluated thoroughly to ensure that the business or the applications' objectives are achieved. Certain metrics can be used for the model evaluation such as accuracy, recall, F1-score, precision, and others.

**Deployment:**

The deployment phase requirements vary from project to project. It can be as simple as creating a report, developing interactive visualization, or making the machine learning model available in the production environment. In this environment, the customers or end-users can utilize the model in different ways such as API, website, or so on.

For this week, you will required to submit the following:

1. Project Title\*
2. A description of the problem and a discussion of the background. (15 marks)
   1. Clearly define a problem or an idea of your choice. Remember that data science problems always target an audience and are meant to help a group of stakeholders solve a problem, so make sure that you explicitly describe your audience and why they would care about your problem.
   2. This submission will eventually become your **Introduction/Business Problem** section in your final report. So I recommend that you push the report (having your Introduction/Business Problem section only for now) to your Github repository and submit a link to it.
   3. Business Understanding: The initial phase is to understand the project's objective from the business or application perspective. Then, you need to translate this knowledge into a machine learning problem with a preliminary plan to achieve the objectives.
3. A description of the data and how it will be used to solve the problem. (15 marks)
   1. Describe the data that you will be using to solve the problem or execute your idea. So make sure that you provide adequate explanation and discussion, with examples, of the data that you will be using.
   2. This submission will eventually become your **Data** section in your final report. So I recommend that you push the report (having your **Data** section) to your Github repository and submit a link to it.
   3. Data understanding: In this phase, you need to collect or extract the dataset from various sources such as csv file or SQL database. Then, you need to determine the attributes (columns) that you will use to train your machine learning model. Also, you will assess the condition of chosen attributes by looking for trends, certain patterns, skewed information, correlations, and so on.
   4. Data Preparation: The data preparation includes all the required activities to construct the final dataset which will be fed into the modeling tools. Data preparation can be performed multiple times and it includes balancing the labeled data, transformation, filling missing data, and cleaning the dataset.

Accident Severity Prediction

The focus of this project is the development of a model that can predict the severity of traffic accidents, given weather and road conditions.

The immediate stakeholders of the project are public safety officials hoping to warn drivers of the likely severity of an accident in order to get them to drive more safely or change their travel plans.

At this stage the model is envisioned to be a supervised machine learning model such as SVM or Decision Tree. Multiple algorithms will be tested to determine which provides the greatest predictive accuracy.

The data to be used for the testing and development of the predictive model will be the shared dataset.

The dataset is rich, with more than 150k observations and many attributes. Beyond the weather and road conditions, there are other variables which seem likely to contribute to the severity of an accident such as light conditions, whether the driver was speeding, distracted, or under-the-influence of a substance.

Key to the development of the model is the fact that the severity label indicates the severity of the observed accidents as either fatal or property damage.

Data preparation will require the handling of missing data, as well as re-structuring certain attributes such as weather conditions. As an example, weather conditions exists in the dataset as a single attribute with several categories. This will be re-structured using indicator variables in order to correlate each separate weather condition with the severity of an accident. The same will be done with other categorical variables. The dataset is also unbalanced with regards to the labeled observation. The dataset will be balanced in order to remove a source of bias in the predictive model.